Capstone 2:

Detecting Deceptive Hotel Reviews

Through Topic Modeling and Machine Learning

SpringBoard Data Science Career Program

Chris Gian

github.com/chrisgian

Roadmap

- Motivation
- Data
- Data Preparation
- Topic Modeling
- Classification and Tuning
- Results
- Future steps

Motivation (1)

Deceptive text-based content plagues many industries and aspects of our information-rich lives.

1. Yelp Fake reviews
   - Increase a business' popularity in rankings
2. Spam
   - Unwanted messages show up on cell phones, emails, and social media
3. Fake news
   - Biased and false content published under the guise of unbiased and legitimate news

Motivation (2)

Yelp's fake review problem

Daniel Roberts
Sep 26, 2013

FORTUNE -- On Monday, New York State Attorney General Eric Schneiderman announced the conclusion of "Operation Clean Turf," a yearlong sting that caught 19 different companies, most of them SEO (search engine optimization) or reputation management firms for hire, that were writing fake reviews for small businesses that paid them. The NYAG fined the businesses in varying amounts that total $350,000. (As part of the operation, people from the NYAG's office even posed as owners of a Brooklyn yogurt shop.) In the press release, Schneiderman says that the investigation "tells us that we should approach online reviews with caution" and calls the process of posting fake reviews online, "the 21st century's version of false advertising."

Motivation (3) Approach

This capstone project focuses on using **Natural Language Processing** and **Machine Learning** methods to identify deceptive content.

Motivation (4) Hypothetical Client

A hotel booking aggregator is trying to improve the quality and reliability of its hotel booking information for its users.

- It has noticed complaints after users voiced concerns that some reviews may be fake.
- To alleviate this, the client seeks to build a predictive model to flag whether reviews are genuine or fake.
- By the end of this project, the client will be able to use this algorithm as part of their fake review deterrence strategy.

Data (1) Can you tell the difference?

	Paid Reviewer	Real Reviewer
Positive	the experince at the hard rock hotel in chicago was fantastic,i will rate them a 6 out of 5. they have wonderful service and great staff and the view is just wonderful.	I recently stayed at the Hard Rock Hotel in Chicago, Il. From the start, the experience was bad. The room was filthy, there were no towels, and the front desk did nothing to rectify the situation. I will never stay there again. I could not have been more dissatisfied.
Negative	The Swissotel Chicago is a very mediocre hotel, the service is always poor, and the room service food always comes cold, unless it's supposed to be cold than it comes warm. I would rather stay at a super 8 than this place again.	I travel often for business and this hotel ranks very low on my list. The room had such a strong odor of smoke, it gave me a headache (and I used to be a smoker)! The room service was mediocre and extremely expensive. The hotel is disturbingly huge. Very difficult to navigate your way around it. I waited on hold for twenty minutes to ask a concierge where a pharmacy was located. She curtly gave me cross-street names and hung up. Uhhhhh, how do I know which direction to go in??? Stains on the carpet in the room. A big gauge in the wall, where maybe a thermostat once was??? The shower is decent. All in all, for the price they charge, NO THANKS!!! I'm just glad my company is footing the bill.

Data (2) Descriptives

Top 5 and Bottom 5 Records

	deceptive	hotel	polarity	source	text
0	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway with family ...
1	truthful	hyatt	positive	TripAdvisor	Triple A rate with upgrade to view room was le...
2	truthful	hyatt	positive	TripAdvisor	This comes a little late as I'm finally catchi...
3	truthful	omni	positive	TripAdvisor	The Omni Chicago really delivers on all fronts...
4	truthful	hyatt	positive	TripAdvisor	I asked for a high floor away from the elevato...
1595	deceptive	intercontinental	negative	MTurk	Problems started when I booked the InterContin...
1596	deceptive	amalfi	negative	MTurk	The Amalfi Hotel has a beautiful website and i...
1597	deceptive	intercontinental	negative	MTurk	The Intercontinental Chicago Magnificent Mile ...
1598	deceptive	palmer	negative	MTurk	The Palmer House Hilton, while it looks good i...
1599	deceptive	amalfi	negative	MTurk	As a former Chicagoan, I'm appalled at the Ama...

Positive and Negative Sentiment by Source

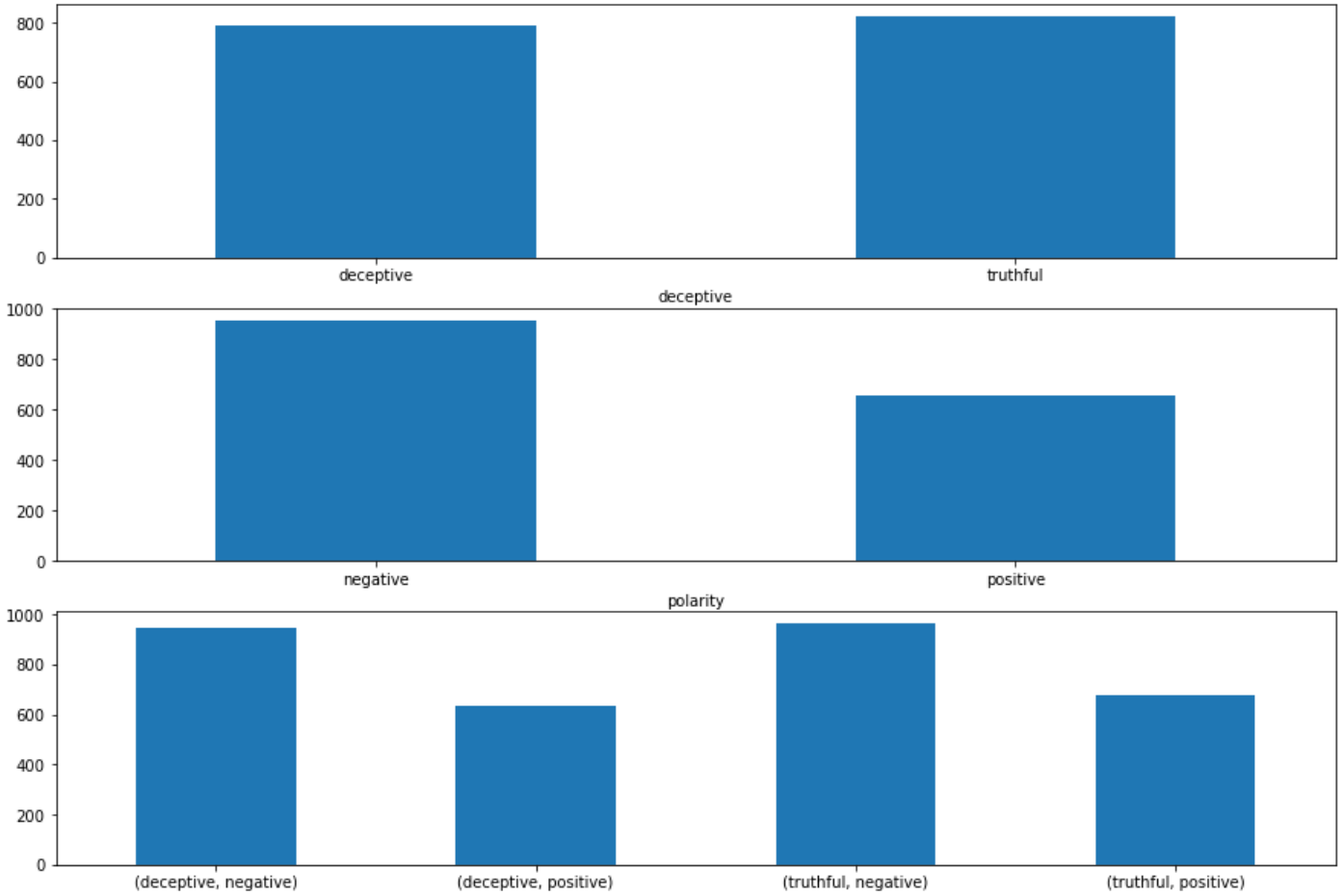
	hotel				text			
deceptive	**deceptive**		**truthful**		**deceptive**		**truthful**	
polarity	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**
source								
MTurk	400.0	400.0	0.0	0.0	400.0	400.0	0.0	0.0
TripAdvisor	0.0	0.0	0.0	400.0	0.0	0.0	0.0	400.0
Web	0.0	0.0	400.0	0.0	0.0	0.0	400.0	0.0

Count of records by Deceptive, Sentiment, and Hotel

	source				text			
deceptive	**deceptive**		**truthful**		**deceptive**		**truthful**	
polarity	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**
hotel								
affinia	20	20	20	20	20	20	20	20
allegro	20	20	20	20	20	20	20	20
amalfi	20	20	20	20	20	20	20	20
ambassador	20	20	20	20	20	20	20	20
conrad	20	20	20	20	20	20	20	20
fairmont	20	20	20	20	20	20	20	20
hardrock	20	20	20	20	20	20	20	20
hilton	20	20	20	20	20	20	20	20
homewood	20	20	20	20	20	20	20	20
hyatt	20	20	20	20	20	20	20	20
intercontinental	20	20	20	20	20	20	20	20
james	20	20	20	20	20	20	20	20
knickerbocker	20	20	20	20	20	20	20	20
monaco	20	20	20	20	20	20	20	20
omni	20	20	20	20	20	20	20	20
palmer	20	20	20	20	20	20	20	20
sheraton	20	20	20	20	20	20	20	20
sofitel	20	20	20	20	20	20	20	20
swissotel	20	20	20	20	20	20	20	20
talbott	20	20	20	20	20	20	20	20

Data (3) Data Exploration

Mean Length of Words Across Categories



From the results of hypotheses tests we can see that:

- There is a statistically significant difference between positive and negative average length of words (p-value of about 0)
- There is a statistically significant difference between trip advisor and non trip advisor sources (p-value of about 0)
- There is a no detectable statistical significance when it comes to deceptive and genuine reviews. (p-value = .2)

Data Preparation: Tokenizing

Sentence → Tokens → Parts of Speech

	deceptive	hotel	polarity	source	text	tokens	tokens_stopwords	lemma	pos
0	truthful	conrad	positive	TripAdvisor	We stayed for a one night getaway with family ...	[We, stayed, for, a, one, night, getaway, with...	[We, stayed, one, night, getaway, family, thur...	[-PRON-, stay, for, a, one, night, getaway, wi...	[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...

Parts of Speech → <u>4 Parts of Speech Variables</u>

pos	pron_ct	noun_ct	punct_ct	verb_ct
[PRON, VERB, ADP, DET, NUM, NOUN, NOUN, ADP, N...	5	30	12	14

Data Preparation: Categorical Data

Categorical Data → Dummy Variables (22 Variables)

hotel	polarity	source
conrad	positive	TripAdvisor
hyatt	positive	TripAdvisor
hyatt	positive	TripAdvisor
omni	positive	TripAdvisor
hyatt	positive	TripAdvisor

hotel_palmer	hotel_sheraton	hotel_sofitel	hotel_swissotel	hotel_talbott	polarity_negative	polarity_positive
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1

Topic Modeling

- Use Latent Semantic Indexing
  - Data Reduction Technique based on “Singular Value Decomposition”
  - Reduces Tokenized Table for each of 1600 review into 300 variables
    - Think of each variable as a “topic” for example:
      - “An amazing Family Vacation” Topic
      - “Honey moon gone awry review” Topic

Classification

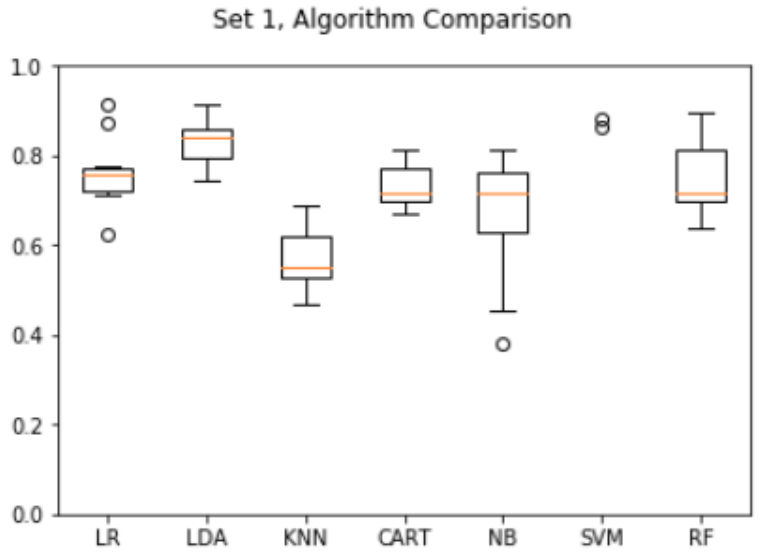
- 6 Classification Models
  - Logistic Regression
  - Linear Discriminant Analysis (LDA)
  - K Nearest Neighbor Classification
  - Decision Trees
  - Naive Bayes (NB)
  - Support Vector Classifier (SVM)
  - Random Forest (RF)

- Three sets of Features:
  - X1 = Topics Only (300 Variables)
  - X2 = Topics + Parts of Speech Metrics (300 + 4)
  - X3 = Topics + Parts of Speech Metrics + Dummy Variables (300 + 4 + 22)

Model Selection (1)

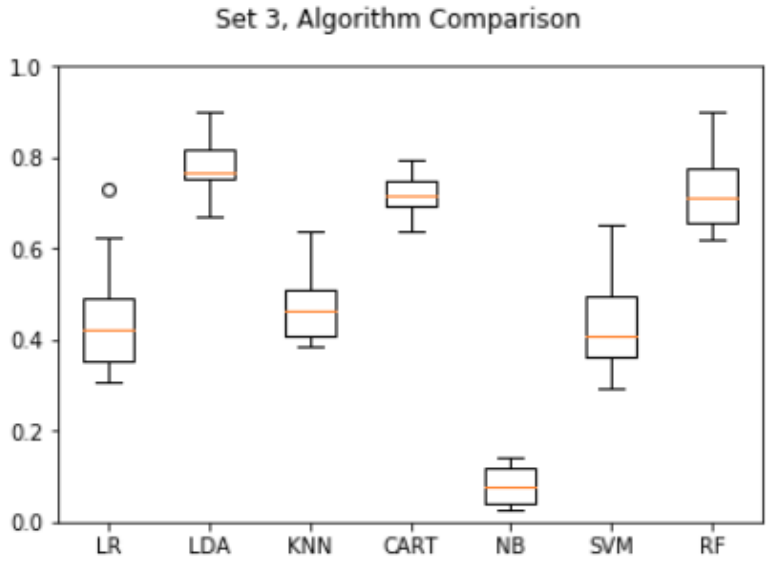
Topics Only

```
LR: 0.761875 (0.077533)
LDA: 0.828750 (0.048750)
KNN: 0.566250 (0.069832)
CART: 0.730000 (0.049117)
NB: 0.667500 (0.136725)
SVM: 0.174375 (0.348775)
RF: 0.748125 (0.080722)
```



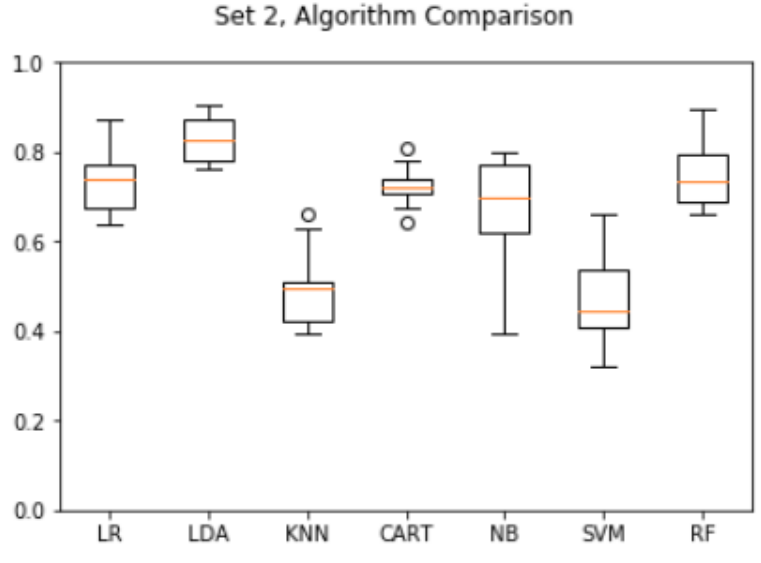
Previous + Parts of Speech Metrics

```
LR: 0.451875 (0.129845)
LDA: 0.780625 (0.062152)
KNN: 0.481875 (0.082701)
CART: 0.720000 (0.045569)
NB: 0.080625 (0.042246)
SVM: 0.436875 (0.111329)
RF: 0.728125 (0.090063)
```



Previous + Dummy Variables

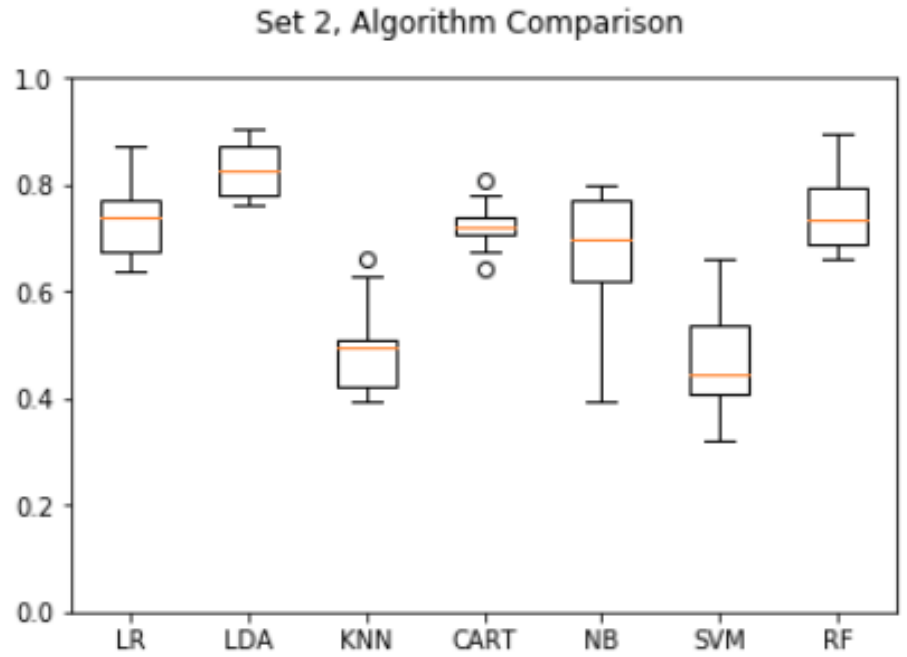
```
LR: 0.736250 (0.072769)
LDA: 0.830000 (0.050683)
KNN: 0.496875 (0.086388)
CART: 0.724375 (0.044586)
NB: 0.659375 (0.135734)
SVM: 0.468125 (0.107654)
RF: 0.750625 (0.076416)
```



LDA performs the best.

Model Selection (2)

```
LR: 0.736250 (0.072769)
LDA: 0.830000 (0.050683)
KNN: 0.496875 (0.086388)
CART: 0.724375 (0.044586)
NB: 0.659375 (0.135734)
SVM: 0.468125 (0.107654)
RF: 0.750625 (0.076416)
```



- LDA has no hyper parameters to be tuned
- Results might be improved by using Random forest with tuned parameters given that without tuning, the model performed reasonably well

Optimizing Random Forest

- Number of Trees to grow: 1000
- Depth of Trees: [15,25,50,75],
- Determine node splits by using Gini Entropy
- Minimum Number of Samples in each node: [50,100]
- Using: 10-Fold Cross Validation

Results (1)

- Best Accuracy: 84.6%
  - 1.6 % improvement compared to LDA
  - + 20% improvement compared to untuned Random Forest model
- Best Model:
  - Max Tree Depth = 50
  - Minimum Samples per Node = 50

Results (2)

- Mean Decrease in Impurity (MDI), we can determine feature importance.
- The figure to the right shows what is most important in determining Deceptive vs Genuine reviews
  - Topic Vectors and 3 are most important
  - The use of punctuation is important

	Importance	Std
2	0.266521	0.281358
6	0.051124	0.081435
3	0.047117	0.076333
4	0.035424	0.060714
12	0.024632	0.047719
15	0.024012	0.047223
punct_ct	0.019903	0.043599

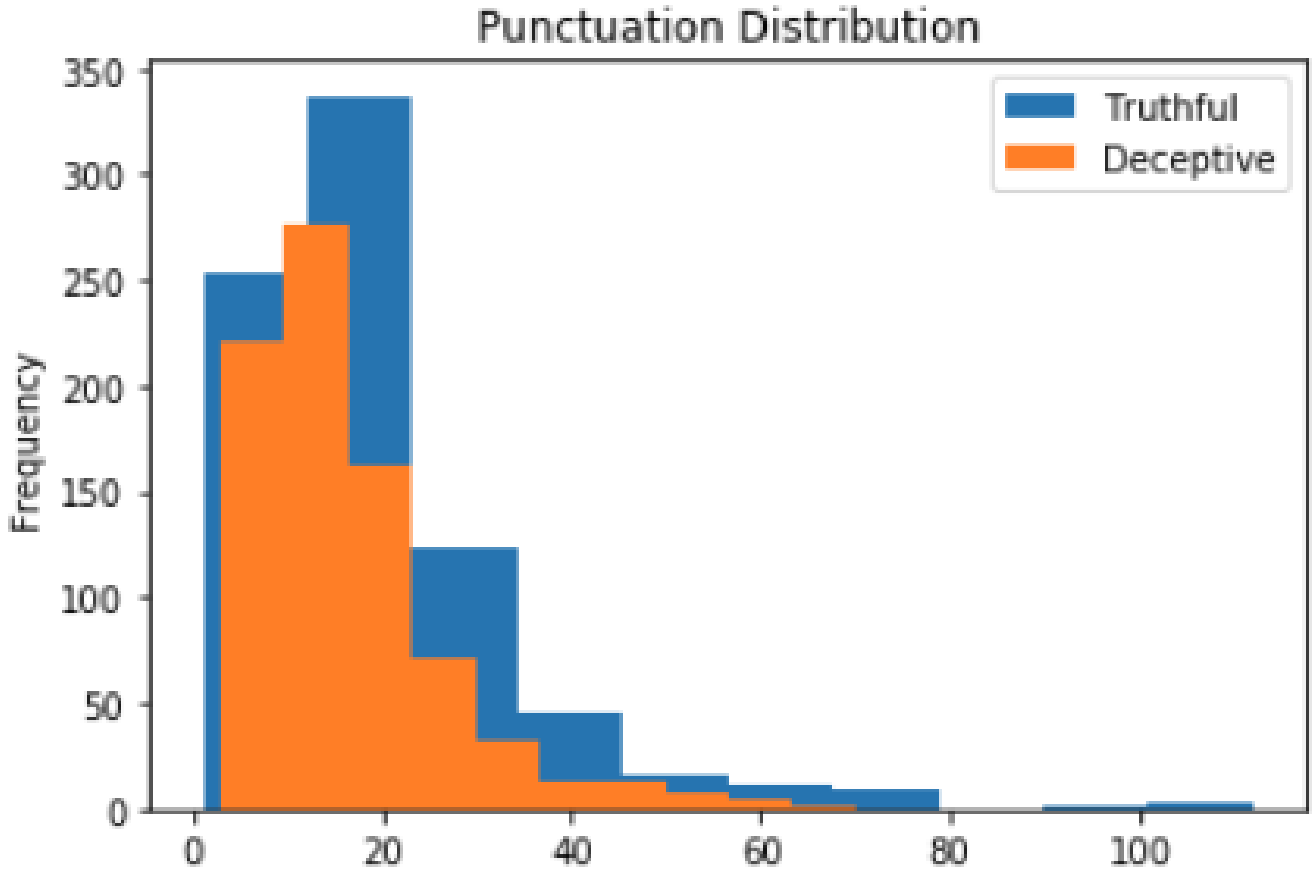
Results (3)

Looking into each topic vector, we see what types of words impact deceptiveness of a review

	Word	Weight	topic
0	'	0.220317	2
1	great	-0.174643	2
2	comfortable	-0.129004	2
3	beautiful	-0.116429	2
4	enjoyed	-0.113951	2
5	definitely	-0.112457	2
6	When	0.110262	2
7	wonderful	-0.109365	2
8	called	0.104809	2
9	helpful	-0.104119	2
0	)	0.231953	3
1	(	0.222104	3
2	...	0.209163	3
3	$	0.204930	3
4	Hotel	-0.145028	3
5	:	0.138617	3
6	-	0.128757	3
7	Michigan	0.122856	3
8	'	0.114440	3
9	husband	-0.102649	3

Results (3)

Descriptively, there is a noticeable difference in punctuation count between truthful and deceptive reviews.



Future steps

- The results above suggest that the model performs fairly well under Random Forest, yet here are some areas of improvement and/or further areas of research.
  - The number of features is large compared to the number of observations
    - Potentially, tune LSI, reduce the Term Frequency-Inverse Document Frequency to a smaller number of topics
  - Word misspellings in feature engineering. Train the models with a misspelling indicator, as this could be a feature that might be important.
  - Look at this data across time -- there could be difference in detecting spam based on information around the time period -- for example, will adding month pick up information about holidays and the holiday vacation experience that could help detect fake reviews.
  - Would having transaction data on each review show interesting results -- for example, the time of each post, and location of the IP address that sent the post.